Final Report: Jan 1, 2004 to Dec 31, 2006
"Classification, Clustering and Dimensionality Reduction",
ONR Award No. N000140410183
Anil K. Jain
Michigan State University
July 8, 2008

The proposal "Classification, Clustering and Dimensionality Reduction" addressed some important issues that remain unsolved in pattern recognition, data mining and machine learning. The key objective of the proposal was to investigate the following important problems in pattern recognition: (i) combination of clustering algorithms, and (ii) dimensionality reduction.

Our work has resulted in solutions to these problems, which we believe have advanced the state of the art in pattern recognition, data mining and machine learning. A brief summary of the accomplishments is provided below, along with the resulting publications.

1. Incremental ISOMAP

Understanding the structure of multidimensional patterns, especially in unsupervised case, is of fundamental importance in data mining, pattern recognition and machine learning. Several algorithms have been proposed to analyze the structure of high dimensional data based on the notion of manifold learning. These algorithms have been used to extract the intrinsic characteristics of different types of high dimensional data by performing nonlinear dimensionality reduction. Most of these algorithms operate in a *batch* mode and cannot be efficiently applied when data are collected sequentially. In this work, we described an incremental version of ISOMAP, one of the key manifold learning algorithms. Our experiments on synthetic data as well as real world images demonstrate that our modified algorithm can maintain an accurate low-dimensional representation of the data in an efficient manner.

M. H. Law, A. K. Jain. "Incremental Nonlinear Dimensionality Reduction by Manifold Learning", *IEEE Transactions of Pattern Analysis and Machine Intelligence*. vol. 28, no. 3, pp. 377 - 391, March 2006.

2. Combination of clustering algorithms

We explored the idea of evidence accumulation (EAC) for combining the results of multiple clusterings on the same data. First, a clustering ensemble - a set of object partitions, is produced. Given a data set (n patterns in d dimensions), different ways of producing data partitions are: 1) applying different clustering algorithms and 2) applying the same clustering algorithm with different values of parameters or initializations.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
7/19/2008	Final	1/1/04-12/31/06

4. TITLE AND SUBTITLE
Classification, Clustering and Dimensionality reduction

5a. CONTRACT NUMBER
N00014-04-1-0183

5b. GRANT NUMBER
N00014-04-1-0183

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Anil K. Jain

5d. PROJECT NUMBER
06PR01012-05

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Michigan State University
Department of Computer Science
Michigan State University
East Lansing, MI 48824

8. PERFORMING ORGANIZATION REPORT NUMBER
MSU-ONR-06-Final

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
875 North Randolph Street
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT
"Approved for Public Release; distribution is Unlimited".

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The primary goal of pattern recognition is supervised or unsupervised classification. Among the various frameworks in which pattern recognition has been traditionally formulated, the statistical approach has been most intensively studied and used in practice. The design of a recognition system requires careful attention to the following issues: feature extraction and selection, cluster analysis, and classifier design and learning. In spite of almost fifty years of research and development in this field, the general problem of recognizing complex patterns with arbitrary orientation, location, and scale remains unsolved. New and emerging applications, such as data mining, web searching, retrieval of multimedia data, face recognition and cursive handwriting recognition, require robust and efficient pattern recognition techniques. The objective of this research proposal is to investigate the following important problems in pattern recognition: (i) classifier evaluation, (ii) one-class classification, (iii) combination of clustering algorithms, and (iv) dimensionality reduction. Solution to these problems will advance the state-of-the-art in pattern recognition, data mining and machine learning. These advances will also be useful to a number of pattern recognition and data mining applications of interest to the Navy.

15. SUBJECT TERMS
Pattern recognition, data clustering, feature extraction

16. SECURITY CLASSIFICATION OF:	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON

Further, combinations of different data representations (feature spaces) and clustering algorithms can also provide a multitude of significantly different data partitionings. We proposed a simple framework for extracting a consistent clustering, given the various partitions in a clustering ensemble. According to the EAC concept, each partition is viewed as an independent evidence of data organization, individual data partitions being combined, based on a voting mechanism, to generate a new n x n similarity matrix between the n patterns. The final data partition of the n patterns is obtained by applying a hierarchical agglomerative clustering algorithm on this matrix. We have developed a theoretical framework for the analysis of the proposed clustering combination strategy and its evaluation, based on the concept of mutual information between data partitions. Stability of the results is evaluated using bootstrapping techniques. A detailed discussion of an evidence accumulation-based clustering algorithm, using a split and merge strategy based on the k-means clustering algorithm, is presented. Experimental results of the proposed method on several synthetic and real data sets are compared with other combination strategies, and with individual clustering results produced by well-known clustering algorithms.

A. Fred, A.K. Jain. Combining Multiple Clustering Using Evidence Accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, number 6, pp. 835-850, 2005.

B. A. Topchy, A.K. Jain, W. Punch, "Clustering Ensembles: Models of Consensus and Weak Partitions", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol., 27, No. 12, pp. 1866-1881, Dec 2005

C. A. Fred and A.K. Jain, "Learning Pairwise Similarity for Data Clustering", in Proc. of International Conference on Pattern Recognition (ICPR), Vol. 1, pp. 925 – 928, Hong Kong, August, 2006. ***Received the Best Paper Award.***

D. 19. A. K. Jain, P. K. Mallapragada and M. Law, "Bayesian Feedback in Data Clustering", in Proc. of International Conference on Pattern Recognition (ICPR), Vol. 3, pp. 374-378, Hong Kong, August, 2006.

3. Feature selection and Mixture fitting

Clustering is a common unsupervised learning technique used to discover group structure in a set of data. While there exist many algorithms for clustering, the important issue of feature selection, that is, what attributes of the data should be used by the clustering algorithms, is rarely touched upon. Feature selection for clustering is difficult because, unlike in supervised learning, there are no class labels for the data and, thus, no obvious criteria to guide the search. Another important problem in clustering is the determination of the number of clusters, which clearly impacts and is influenced by the feature selection issue. In this paper, we propose the concept of feature saliency and introduce an expectation-maximization (EM) algorithm to estimate it, in the context of mixture-based clustering. Due to the introduction of a minimum message length model selection criterion, the saliency of irrelevant features is driven toward zero, which corresponds to

performing feature selection. The criterion and algorithm are then extended to simultaneously estimate the feature saliencies and the number of clusters.

A. M. Farmer and A.K. Jain, "Smart Automotive Airbags: Occupant Classification and Tracking", *IEEE Trans. Vehicular Technology*, Vol. 52, No. 1, pp. 60-80, Jan. 2007

B. M.E. Farmer, S.B. Farmer and A.K. Jain, "Non-parametric Feature Selection for Image-based Airbag Suppression", Proc. Applied Statistics Conf., Slovenia, Sept. 18-21, 2005.

C. M. Law, M. A. T. Figueiredo, A. K. Jain. "Simultaneous Feature Selection and Clustering Using Mixture Models", *IEEE Transactions of Pattern Analysis and Machine Intelligence.* vol. 26, no. 9, pp. 1154- 1166, September 2004.